

---

# Improved recognition by combining different features and different systems

Dan Ellis  
International Computer Science Institute  
Berkeley CA  
dpwe@icsi.berkeley.edu

## Outline

- 1 The power of combination**
- 2 Different ways to combine**
- 3 Examples & results**
- 4 Conclusions**



---

---

# 1

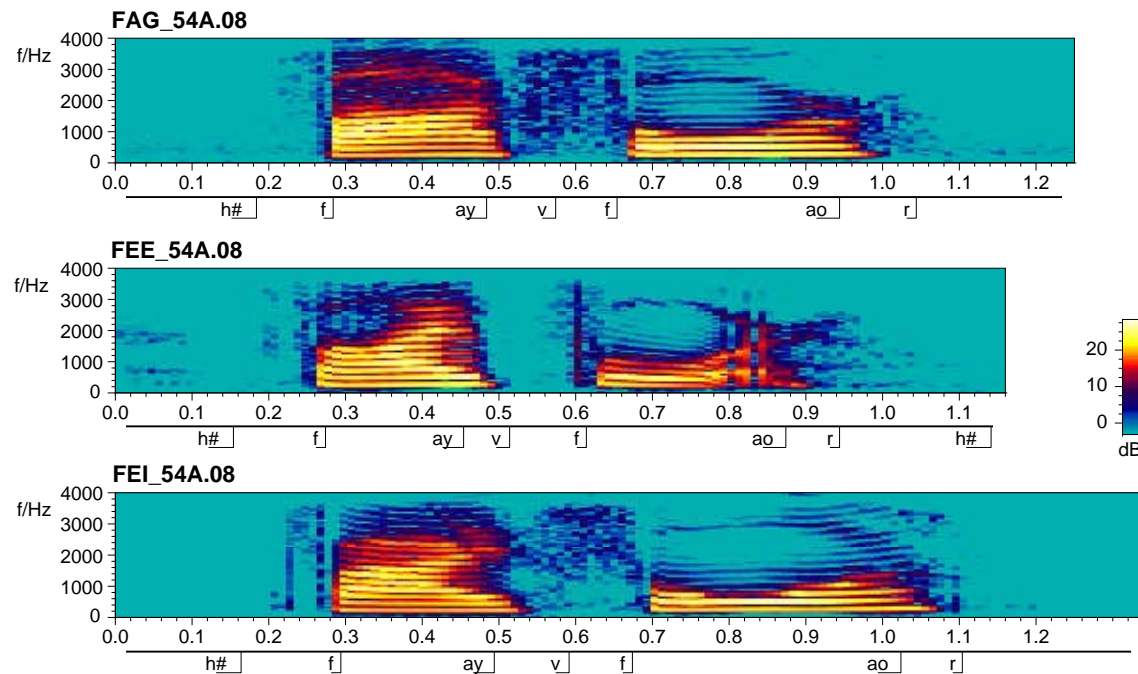
## The power of combination

- **Combination is a general approach in statistics**
  - several models → several estimates
  - if 'correct' parts more consistent than 'wrong' parts...  
→ averaging reduces error
- **Intuition: choose 'right' model for each case**
  - .. need to know when each is 'right'
  - .. need models that are 'right' at different times
- **Continuum of combination schemes**
  - conservative, ignorant of models
  - domain-specific, trained on models



# Relevant aspects of the speech signal

- **Speech is highly redundant**
  - intelligible despite large distortions
  - multiple cues for each phoneme
- **Speech is very variable**
  - redundancy leaves room for variability
  - speakers can use different subsets of cues



---

# Combinations for speech recognition

- **Speech recognition abounds with different models**
  - different feature processing, statistical models, search techniques ...
- **Redundancy in signal**
  - many different ways to estimate, making different kinds of errors
- **General combination is easier than figuring out what is good in each estimator**
  - training data & training process usually the limiting factor



---

---

# Outline

## 1 The power of combination

## 2 Different ways to combine

- Feature combination
- Posterior combination
- Hypothesis combination
- System hybrids

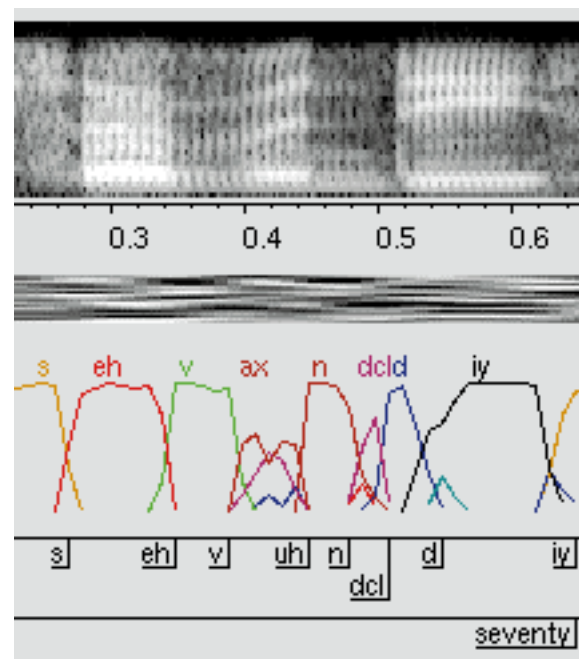
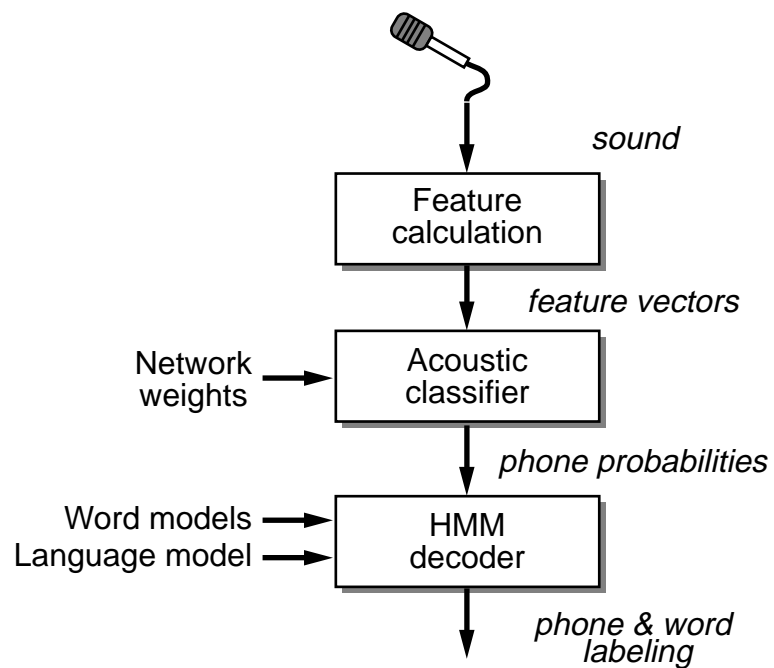
## 3 Examples & results

## 4 Conclusions



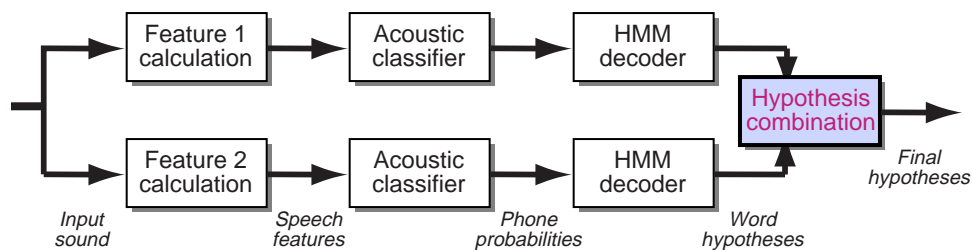
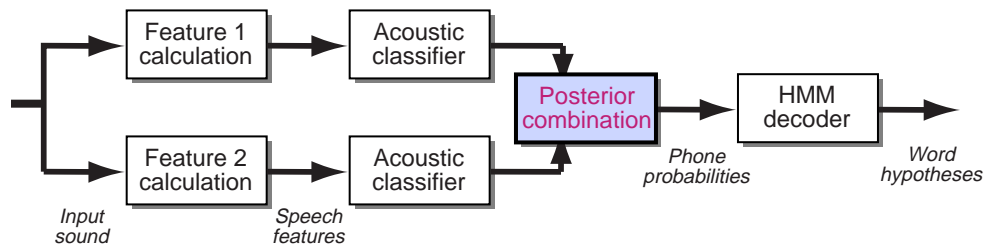
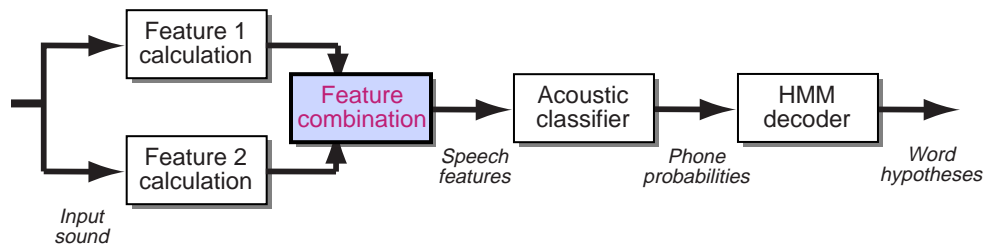
## 2

# Speech recognizer components



# Different ways to combine

- After each stage of the recognizer



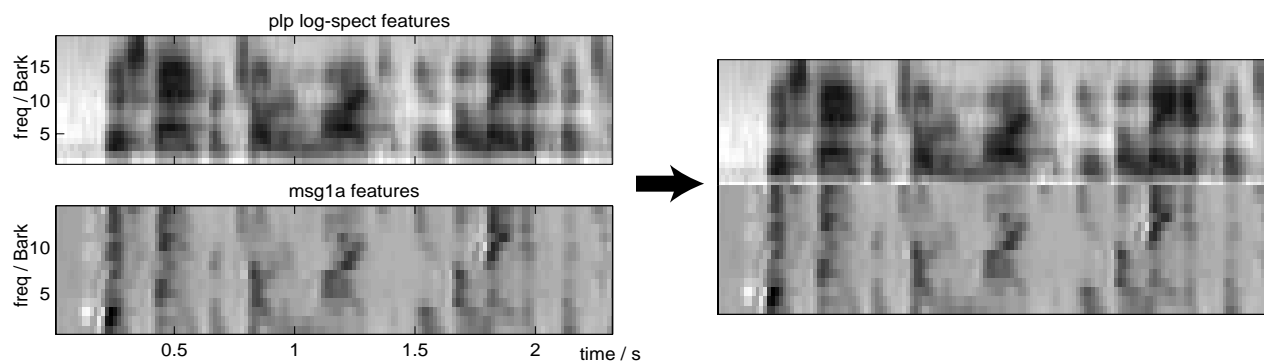
- Lots of other ways...



---

## Feature combination (FC)

- Concatenate different feature vectors
- Train a single acoustic model



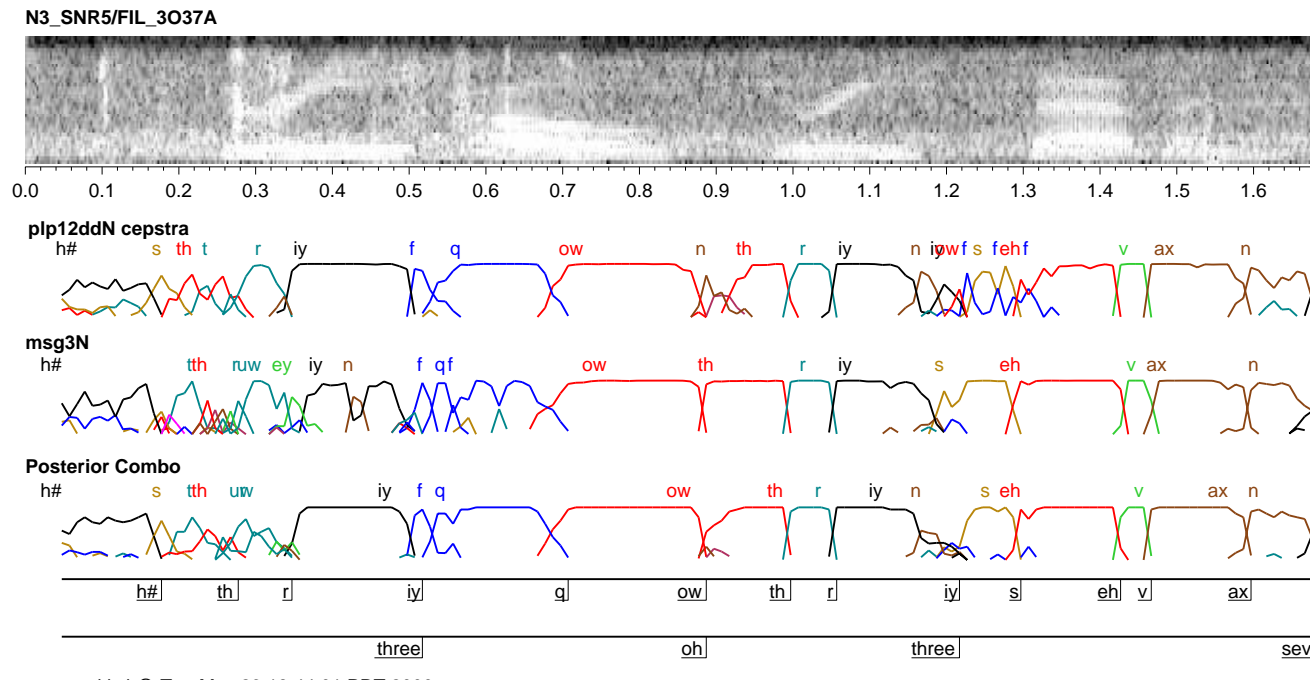
- It helps:

| <i>Features</i> | <i>Avg. WER</i> |
|-----------------|-----------------|
| plp             | 8.9%            |
| msg             | 9.5%            |
| FC plp + msg    | 8.1%            |





# Posterior combination (PC)



- Sometimes better than FC:

| <i>Features</i> | <i>Avg. WER</i> |
|-----------------|-----------------|
| plp             | 8.9%            |
| msg             | 9.5%            |
| PC plp + msg    | 7.1%            |



---

# Hypothesis combination (HC)

(J. Fiscus, NIST)

- **ROVER:**  
Recognizer Output Voting Error Reduction
- Final outputs from several recognizers, each word tagged with confidence
- Align & vote for resulting word sequence

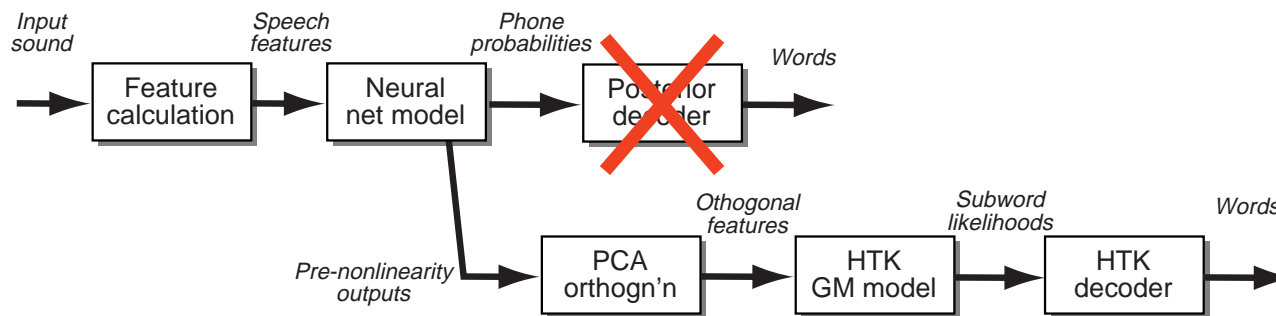
|              |         |     |       |    |     |       |           |    |      |         |           |     |       |
|--------------|---------|-----|-------|----|-----|-------|-----------|----|------|---------|-----------|-----|-------|
| bbn1.ctm     | there's | a   | lot   | of | @   | like  | societies | @  | @    | ruin    | engineers | and | lakes |
| cmu-is11.ctm | there's | the | labs  | @  | @   | like  | societies | @  | for  | women   | engineers | i   | think |
| cu-htk2.ctm  | there's | the | last  | @  | @   | like  | societies | @  | true | of      | engineers | and | like  |
| dragon1.ctm  | was     | @   | alive | @  | the | legal | society   | is | for  | women   | engineers | and | like  |
| sril.ctm     | there's | a   | lot   | of | @   | like  | society's | @  | @    | through | engineers | @   | like  |

- **25% relative improvement over best single Broadcast News system (14.1%→10.6%)**



## Other combinations: 'Tandem' acoustic modeling (with Hermansky et al., OGI)

- **Can we combine with conventional models?**



- **Result: better performance than either alone!**
  - neural net & Gaussian mixture models extract different information from training data

| <i>System-features</i> | <i>Avg. WER</i> |
|------------------------|-----------------|
| HTK-mfcc               | 13.7%           |
| Neural net-mfcc        | 9.3%            |
| Tandem-mfcc            | 7.4%            |



---

# How & what to combine

(with Jeff Bilmes, U. Washington)

- **Combination is good, but...**
  - which streams should we combine?
  - which combination methods to use?
- **Best streams to combine have complementary information**
  - can measure as Conditional Mutual Information,  
 $I(X ; Y \mid C)$
  - low CMI suggests combination potential
- **Choice of combination method depends:**
  - FC for streams with complex interdependence
  - PC makes best use of limited training data
  - HC allows different subword units



---

---

# Outline

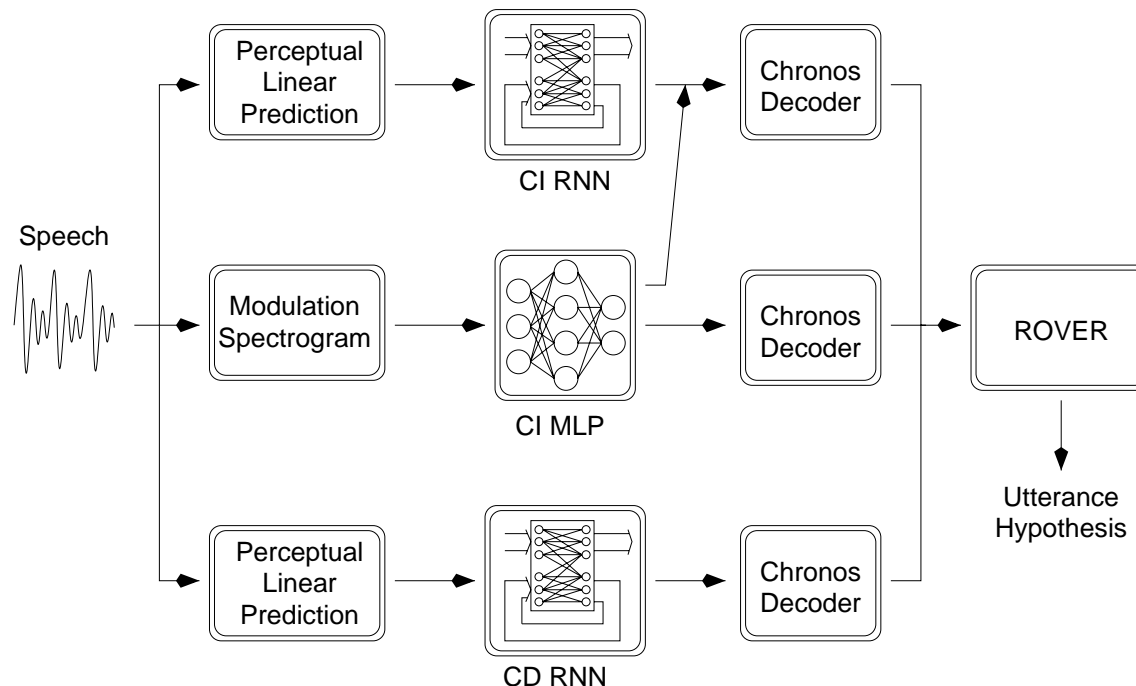
- 1 The power of combination
- 2 Different ways to combine
- 3 Examples & results**
  - The SPRACH system for Broadcast News
  - OGI-ICSI-Qualcomm Aurora recognizer
- 4 Conclusions



# The SPRACH Broadcast News recognizer

(with Cambridge & Sheffield)

- **Multiple feature streams**
- **MLP and RNN models, including PC**
- **Rover for hypothesis combination**



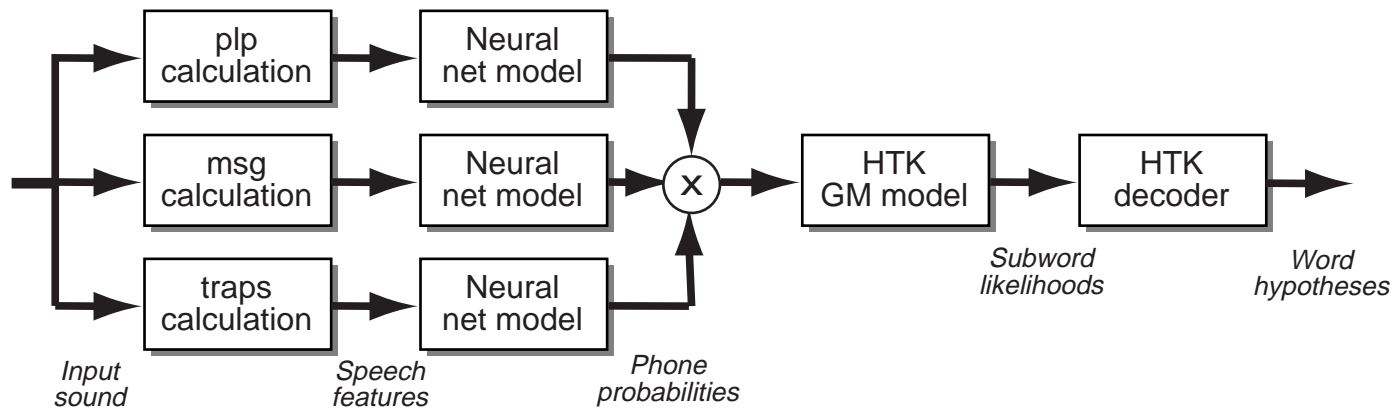
- **20% WER overall (c/w ~27% per stream)**



---

# The OGI-ICSI-Qualcomm system for Aurora noisy digits

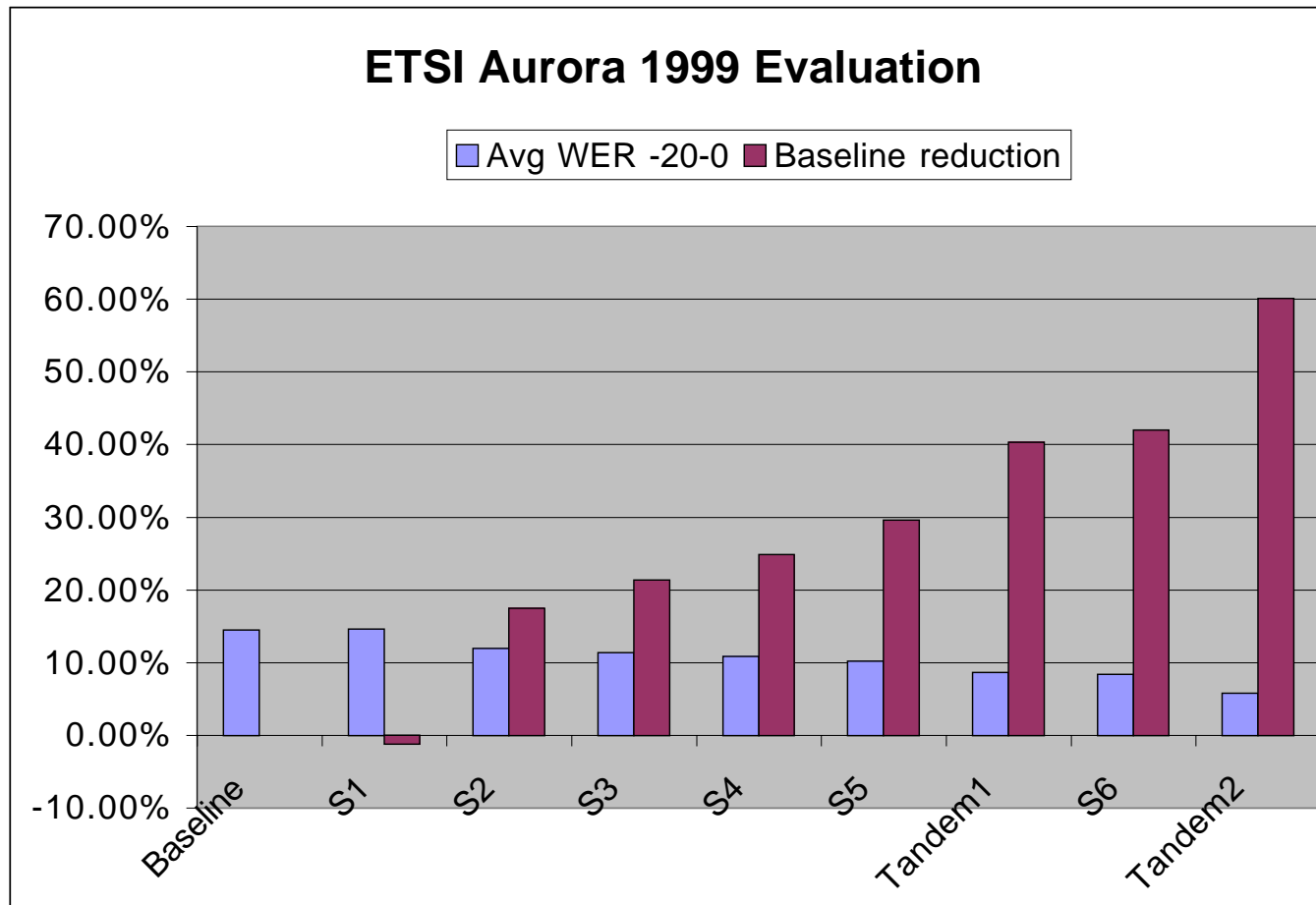
- PC of feature streams
- + Tandem combination of neural nets and Gaussian mixture models:



- 60% reduction in word errors compared to MFCC-HTK baseline



# Aurora evaluation results





---

---

## 4

# Conclusions

- **Combination is a simple way to leverage multiple models**
  - speech recognition has lots of models
  - redundancy in the speech signals allows each model to find different 'information'
- **Lots of ways to combine**
  - after feature extraction, classification, decoding
  - 'tandem' hybrids etc.
- **Significant gains from simple approaches**
  - e.g. 25% relative improvement from posterior averaging
- **Better insight into sources of benefits will lead to even greater gains**

